PHOSIDA 2011: the posttranslational modification database

Florian Gnad[1,2], Jeremy Gunawardena[2] and Matthias Mann[1,*]

[1]Department of Proteomics and Signal Transduction, Max-Planck-Institute for Biochemistry, Am Klopferspitz 18, D-82152 Martinsried, Germany and [2]Department of Systems Biology, Harvard Medical School, 200 Longwood Avenue, Boston, MA 02115, USA

ABSTRACT

The primary purpose of PHOSIDA (http://www.phosida.com) is to manage posttranslational modification sites of various species ranging from bacteria to human. Since its last report, PHOSIDA has grown significantly in size and evolved in scope. It comprises more than 80 000 phosphorylated, N-glycosylated or acetylated sites from nine different species. All sites are obtained from high-resolution mass spectrometric data using the same stringent quality criteria. One of the main distinguishing features of PHOSIDA is the provision of a wide range of analysis tools. PHOSIDA is comprised of three main components: the database environment, the prediction platform and the toolkit section. The database environment integrates and combines high-resolution proteomic data with multiple annotations. High-accuracy species-specific phosphorylation and acetylation site predictors, trained on the modification sites contained in PHOSIDA, allow the in silico determination of modified sites on any protein on the basis of the primary sequence. The toolkit section contains methods that search for sequence motif matches or identify de novo consensus, sequences from large scale data sets.

INTRODUCTION

Many cellular events are controlled by the post-translational modification (PTM) of specific proteins in the proteome. For example, almost all signaling pathways are controlled by reversible phosphorylation, ubiquitination and other PTMs (1,2). In recent years, mass spectrometry (MS)-based proteomics has proven a powerful and generic tool to study these events on a global scale (3). PHOSIDA provides a repository for such modification sites and a systematic approach to protein and site annotation that requires integrating and standardizing data from various sources. It started in 2006, when the Mann laboratory described a generic, quantitative and high-resolution MS technology for the identification and quantitation of phosphorylation sites as a function of stimulus and time (4). Human cells were stimulated with EGF and site specific phosphorylation dynamics were determined and the resulting MS data were recorded in the PHOSIDA database. This study provided a blueprint for many subsequent large scale phosphoproteomic studies in the Mann group (5–7) and the impetus to develop PHOSIDA into a comprehensive and integrative environment. Initially, the main purpose of PHOSIDA was to make the identified phosphorylation data publicly and easily accessible. This explains its original name 'PHOSIDA—the PHOsphorylation SIte DAtabase'. However, with the rapidly increasing number of identified PTM sites from different species, the systematic integration of various annotation data and the provision of multiple analysis tools, PHOSIDA became much more than a mere database. The first extensions including the evolutionary analysis and prediction of human phosphorylation sites were reported in 2007 (8). Since then many additional data sets and features have been integrated into PHOSIDA. The current version manages more than 70 000 phosphorylation sites, and the largest acetylome (9) and N-glycoproteome (10) determined so far. Due to these extensions we now rename PHOSIDA to the 'Posttranslational Modification Site Database'. It contains modification sites of human, mouse, fly, worm and yeast proteins, and is also the most comprehensive repository of prokaryotic phosphoproteomes. To our knowledge, PHOSIDA and Uniprot (11) are the only resources that manage differently modified proteins from such a variety of species. However, Uniprot does not provide the same level of detail on MS and site specific data. Phospho.ELM (12), PhosphoSite (13), HPRD (14) and dbPTM (15) are further comprehensive databases that contain phosphorylation sites from different projects.

Each of these websites has unique features regarding data integration, data representation and annotation. Importantly the identification of all integrated PTM sites in PHOSIDA has been based on high-accuracy mass spectrometry measurements using very strict detection criteria (16). This ensures very small false positive rates, which are not inflated by diverse data sets analyzed by different criteria. Furthermore, the inclusion of quantitative PTM dynamics is a unique feature of PHOSIDA.

The presentation of different annotation data and the provision of analysis tools in PHOSIDA as an integrative platform has recently allowed the large scale evolutionary analysis of phosphorylation in all domains of life (17). Based on integrated phylogenetic relationships, global sequence alignments and structure information, we found that most of the identified eukaryotic phosphoproteins were already present in the earliest forms of life. However, their regulation via phosphorylation evolved after the divergence between single- and multi-cellular species. Even the worm phosphoproteome was found to be very distinct from the phosphoproteomes of higher eukaryotes, which is in concordance with the evolution of the corresponding kinase families. As another example, the high-accuracy predictors and sequence analysis tools have proven helpful in diverse studies (18–20). Here we describe the current version of PHOSIDA, which contains three main components: the integrative database management, the assembly of species-specific modification predictors and the analysis toolkit.

DATABASE ENVIRONMENT

Posttranslational modification site data sets

Initially PHOSIDA contained more than 6000 phosphorylation sites from HeLa cells exposed to growth factor stimulation. This data set presented the largest identified phosphoproteome at the time but this number has been exceeded in many subsequent large-scale studies, with no sign of saturation (Table 1). The current version of PHOSIDA contains an additional 20 000 human phosphorylation sites from kinase enriched samples (21) or of

Table 1. Number of identified posttranslationally modified proteins, peptides and sites

	Proteins	Peptides	Sites
Phosphorylation			
Homo sapiens	8283	23 130	24 262
Mus musculus	9234	24 604	25 085
Drosophila melanogaster	2379	8777	10 043
Caenorhabditis elegans	2373	6926	6780
Saccharomyces cerevisiae	1118	4457	3620
Lactococcus lactis	63	99	73
Bacillus subtilis	78	102	76
Escherichia coli	79	104	81
Halobacterium salinarium	62	100	75
Acetylation			
Homo sapiens	1750	4219	3600
N-Glycosylation			
Mus musculus	2352	8681	6367

quantified dynamics during the cell cycle (5,7). Furthermore about 25 000 mouse phosphorylation sites are managed by PHOSIDA. These sites were derived from liver cells (6), melanoma tissue (22), brain (23) or macrophages (24). The global phosphoproteomes of fly (25), worm (26) and yeast (27) have also been added, so that PHOSIDA covers several representative eukaryotic species. Interestingly, the overlaps between different phosphoproteomes are relatively low on the site level, as demonstrated in the associated studies. This underlines the vast extent of phosphorylation in the cell. It also indicates that the identification of the complete eukaryotic phosphoproteome is far from being achieved. While thousands of serine/threonine and tyrosine phosphorylation sites can be detected in eukaryotic cells, measured prokaryotic phosphoproteomes generally do not comprise more than 100 sites. The relatively low extent of phosphorylation can be observed for both gram-negative and gram-positive bacteria such as *Escherichia coli* (28) and *Bacillus subtilis* (29), respectively. In the third domain of life, the archaeans, the detected serine/threonine and tyrosine phosphoproteome is likewise limited to 75 sites (30). Notably, mitochondria—the eukaryotic organelles with prokaryotic origin—also comprise a relatively sparse phosphoproteome, which classifies them with the prokaryotes rather than with other mammalian organelles (17).

In addition to phosphorylation data, 3600 acetylated lysines (9) and 6367 N-glycosylated asparagines (26) have been uploaded to PHOSIDA. To allow the species-specific retrieval of PTM sites from studies which employed different databases for identification, detected modified peptides and proteins are regularly reassigned to up-to-date database versions.

Another unique feature of PHOSIDA is the uniform quality of the data. Acceptance of all PTM sites was based on high-accuracy mass spectrometry with stringent criteria yielding a very low false positive rate in the whole repository. Additionally, the online application additionally enables retrieval of modified sites from other sources including the Swiss-Prot database, Phospho.ELM and PhosphoSite. The collaboration with Phospho.ELM, another specialized modification site database, in particular, has proven very fruitful, ensuring up-to-date linkage. This could provide a model of PTM exchange akin to the exchanges of mainstream protein and gene databases. The following syntax provides a link directly to the annotation information of any eukaryotic protein of interest: http://141.61.102.18/phosida/index.aspx?query = [*Uniprotaccession number*].

Searching and browsing

For each species users can search for any protein of interest via accession number, gene name, description or sequence. As one of the major improvements, one can now browse for all posttranslationally modified proteins that were identified in a particular experiment or cell type (Figure 1). In addition, a gene ontology filter allows the retrieval of modified proteins that are localized in a certain cellular compartment or have a specified molecular

Figure 1. The new browsing function allows the searching for posttranslationally modified proteins that were identified in a particular experiment or cell type. Furthermore, the gene ontology filter enables users to search for modified proteins with specific cellular localization and molecular function (left panel). Selecting one of the resulting protein entries (middle panel) yields the protein annotation web page (right panel). In addition to general protein information, identified posttranslational modification sites are listed. Clicking on one of the site buttons results in the provision of site-specific information. In the illustrated example, searching for protein kinases that are both phosphorylated and N-glycosylated in the mouse brain and localized in the plasma membrane yields a list of proteins that match the specified search criteria. One of these proteins is the insulin receptor.

function. The gene ontology data were derived from the AMIGO website (31,32). For example, users can browse for all protein kinases that are both phosphorylated and N-glycosylated in the mouse brain and localized at the plasma membrane.

Posttranslational modification site information and integrated annotation data

For each modified protein, the user is presented with features such as description, gene symbol, sequence, accession numbers from various databases and gene ontology annotation. For the latter category the full terms (e.g. 'ATPase activity') with links to the corresponding entry of the gene ontology website are given. In the case of eukaryotic Swiss-Prot annotated proteins motifs, domains, modified sites from other sources and associated literature references with links to the PubMed site are provided. The integration of annotation data has proven very informative—for example, it allows immediate visualization of PTMs that occur in a certain domain. Since the localization of sites within the detected posttranslationally modified peptide is sometimes ambiguous, we had developed a probability based localization score (4). It reflects the chance of each site within the peptide to be posttranslationally modified given its fragmentation spectra. 'Class I sites' are defined by a minimum localization probability of 0.75. If this score is lower than 0.75, the site is enclosed in brackets in PHOSIDA. For each PTM site, the corresponding localization scores, the surrounding sequence, matching sequence motifs and the predicted secondary structure and accessibility are provided (Figure 2 left panel). As a major difference to previous releases, it is shown whether the specified site was detected in a certain experiment or cell type, if applicable

(Figure 2 right panel). Clicking on one of the corresponding buttons yields the listing of the identified corresponding peptides and quantitative data, if available. For the most recent studies the related spectra are shown for additional validation. The indication of the occurrence of PTM sites in cancer cell lines or normal tissues is a striking feature of PHOSIDA. As previously discussed (33) modification states observed in cell lines might not occur in normal cancer tissues and their interpretation might therefore be misleading. The 'help' section lists all sample conditions used in the associated experiments.

The evolutionary section provides information about the phylogenetic relationships between modified proteins and homologs in other species as described (8). Based on global alignments the amino acid conservation of all identified sites in orthologous proteins is displayed. The aligned surrounding sequences demonstrate if matching motifs are conserved. In contrast to the previous version of PHOSIDA, we now use the phylogentic relationships of 36 eukaryotes provided by the Ensembl Compara database (34).

POSTTRANSLATIONAL MODIFICATION SITE PREDICTION

Target serine/threonine/tyrosine sites are generally recognized by kinases and phosphatases through linear sequence patterns (motifs). Analogously, lysine acetylation sites are recognized by acetyltransferases and deacetylases. Various machine learning approaches try to predict phosphorylation sites. For example, Scansite (35) uses a profile method, whereas the prediction system NetPhos (36) is based on neural networks to predict phosphorylation events. NetPhosK (37) aims to

Figure 2. On the site level PHOSIDA provides the surrounding sequence, matching motifs, the predicted secondary structure, the predicted accessibility, the corresponding identified peptides and the identification state in certain cell types (left panel: N-glycosylated asparagine on position 426 of the mouse glutamate receptor 2 subunit) or experiments (right panel: serine on position 1039 of the human EGF receptor).

predict phosphorylation sites along with their corresponding kinase. Each prediction method is unique regarding the underlying machine learning method, input sets used for training and usability. The main advantages of the PHOSIDA predictors are the high quality of the PTM sites used as input sets for training, the species specificity, and the particular effort invested in user-friendliness.

We use our large-scale studies to construct a PTM site predictor based on a support vector machine. Using the integrated high-resolution MS data sets we developed support vector machines to predict phosphorylation and acetylation sites on the basis of the primary sequence. PHOSIDA contains phosphorylation site predictors for yeast, worm, fly, mouse and human. As previously shown (38,39), phosphorylation prediction accuracy can increase by the addition of further features such as structure and conservation. However, in our studies the prediction accuracies were already high on the basis of the surrounding sequences. The addition of further information including structural constraints and conservation yielded only a slight increase in prediction accuracy (8). The species specificity of the predictors proved to be crucial, as the accuracy decreases with input sets from distantly related species (25). For example, the accuracy of identifying yeast phosphosites using the human phosphorylation site predictor is comparatively low. Novel identified sites are continuously used to generate larger species-specific training sets. However, the addition of

further sites results only in a slight increase of the accuracy. Moreover, PHOSIDA provides a mouse acetylation site predictor with 78% precision at 78% recall (40). A recent study has shown that PHOSIDA outperforms other acetylation site predictors including LysAcet (41) and PredMod (42). However, their accuracies might also increase upon training with current large scale high-quality data sets. As more lysine acetylomes of other species are mapped by high-resolution MS, it will be interesting to see if the PHOSIDA predictor is also capable to identify acetylation sites from distantly related organisms.

To predict the occurrence of phosphorylated or acetylated sites on a single protein of interest, one can either insert its protein sequence without further annotation or a sequence entry in FASTA format. Addressing web users' feedback, PHOSIDA now allows the prediction of PTM sites on multiple proteins in FASTA format (Figure 3). Users can set a desired cutoff directly on the precision-recall-curve. Restrictions on the format of input sequences are described in the corresponding help section accessible via the 'question mark' button.

TOOLKIT SECTION

The recently established toolkit section contains various sequence analysis methods. The 'motif matcher' searches for matching motifs in any sequence of interest (Figure 4).
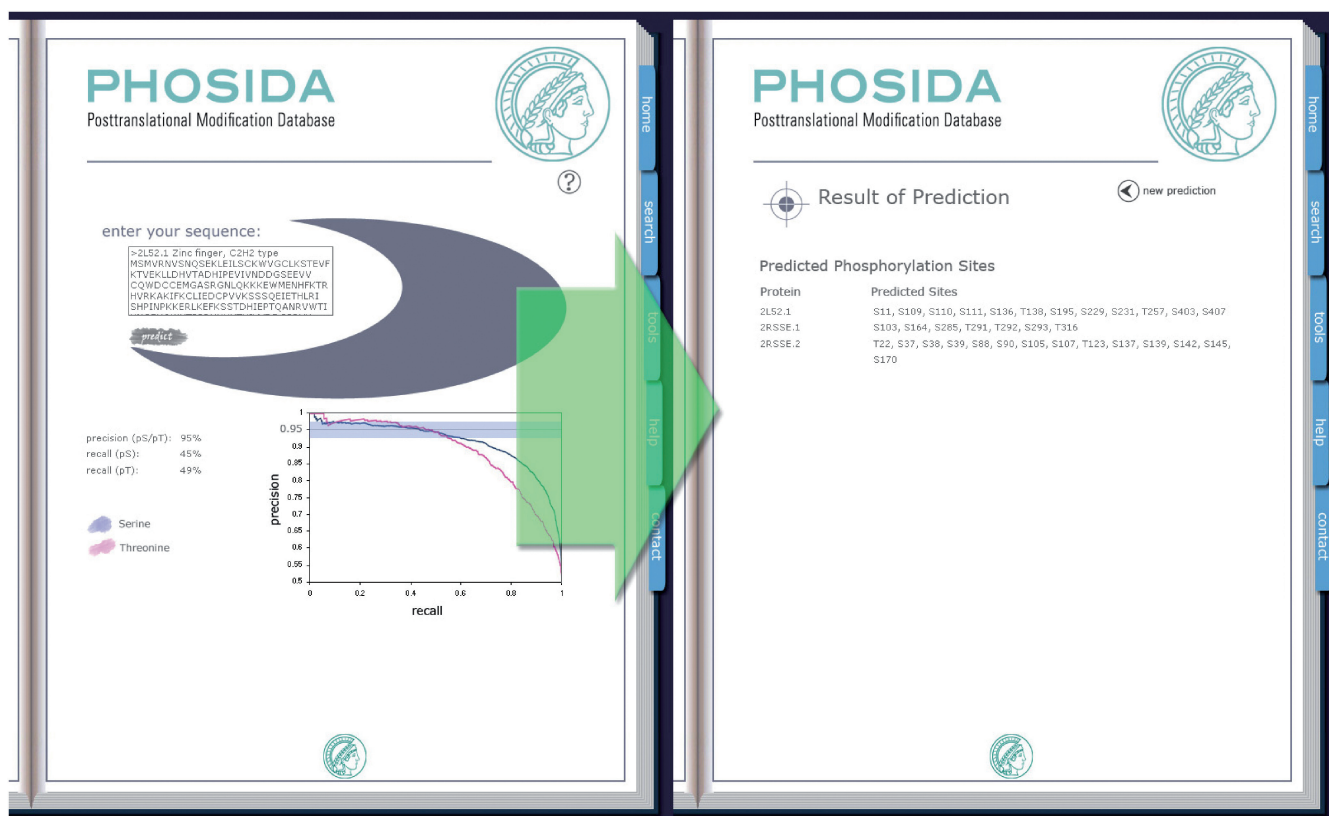
Figure 3. Species-specific phosphorylation or acetylation site predictors allow the *in silico* identification of proteins based on the primary sequence. Users can insert the sequence of a single protein or the sequences of multiple proteins in FASTA format (left panel). Using specified precision recall values, the predicted posttranslational modification sites are listed (right panel).

The underlying repository of annotated sequence motifs contain recognition patterns related to phosphorylation, N-glycosylation and SUMOylation. Alternatively, users can define their own motif and determine matching sites.

Furthermore, we created a 'motif finder' for the *de novo* identification of protein phosphorylation sequence motifs from large scale data sets on the basis of bootstrap statistics. Briefly, sequences surrounding non-phosphorylated serines, threonines and tyrosines are randomly selected from species-specific protein databases in iterative steps. The resulting bootstrap distributions reflect the frequencies of amino acids at certain positions relative to the site. Significantly overrepresented phosphorylation motifs are identified by comparing the position specific amino acid frequencies in the sequences surrounding phosphorylation sites (positive set) with the corresponding calculated bootstrapping distributions (background set). Identified protein sequence motifs are then scanned for matches with annotated kinase motifs. For each derived motif, the corresponding score reflects the difference between the frequency of the given amino acid on the specified position and the mean of the corresponding background bootstrap distribution measured in number of standard deviations of the bootstrap distribution.

Resulting sequence logos visualize the significance of position specific amino acid frequencies from given phosphorylation data sets. Amino acids are displayed in the sequence logo, if their frequency on a given position is higher than the mean of the corresponding background distribution. The height of the amino acid letter is relative to the highest motif identification score.

PHOSIDA currently provides background sets consisting of non-phosphorylated sites of 47 eukaryotes and we intend to add further precalculated background sets to our database in the near future (Figure 5). The background sets are limited to eukaryotic species, as the application to the sparse prokaryotic phosphoproteomes does not yield any significant sequence patterns. The required input sets are phosphorylation sites with their six surrounding residues (to both termini). Phosphosite entries have to be separated via new lines. The input set can be a mixture of phosphoserines, phosphothreonines and phosphotyrosines, and has to contain a minimum of 100 instances for at least one phosphorylated amino acid type. Our online method applies the algorithm to each phosphorylated amino acid (S/T/Y) separately. Consequently, the web user can specify different score and occurrence cutoffs for each phosphorylated amino acid. However, the specified cutoffs have to satisfy minimum values (minimum proportional occurrence of 5% and minimum score of 15). The application to eukaryotic phosphoproteomes shows that many annotated kinase motifs are covered by our *de novo* method. More detailed descriptions of the algorithm and the usage of the
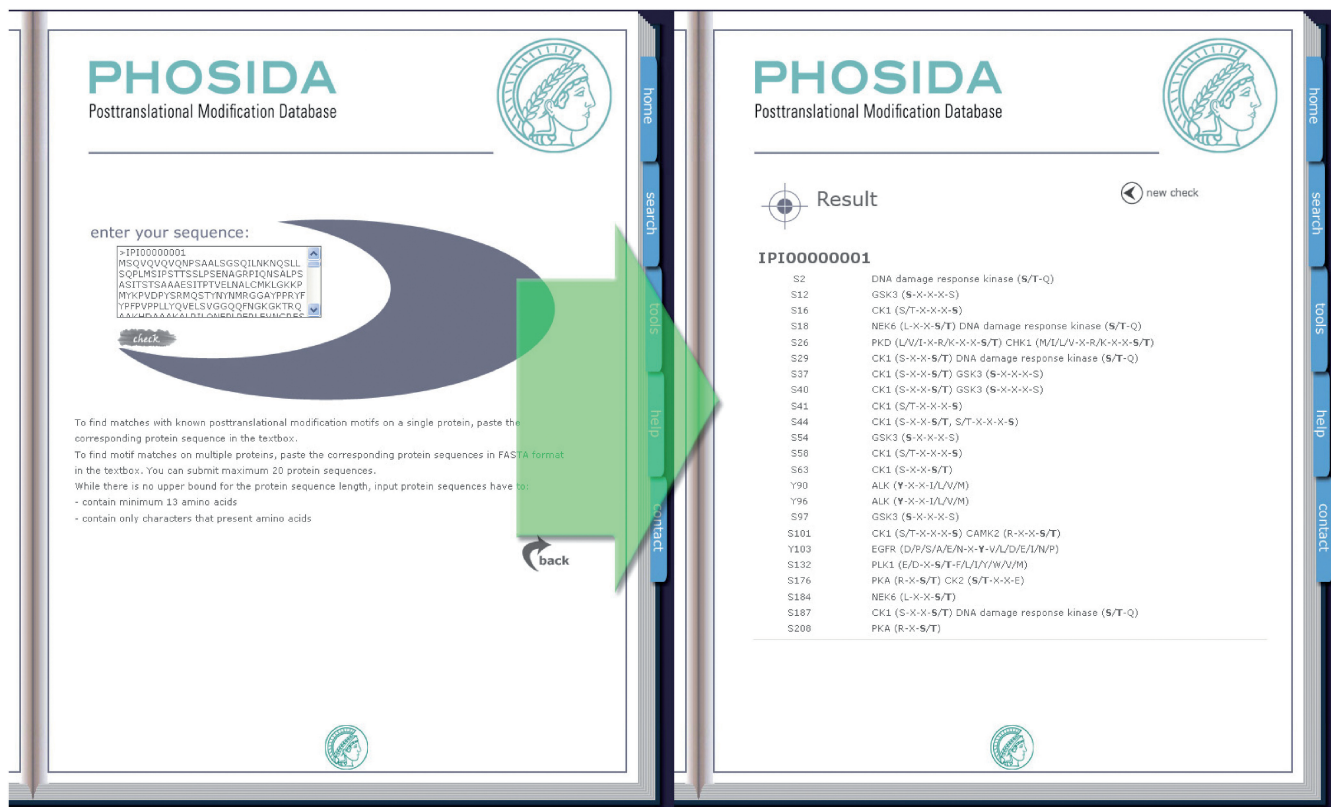
Figure 4. The Motif Matcher searches for sequence matches with annotated motifs including kinase recognition patterns. Users can insert a single sequence or multiple sequences (left panel) to find motif matches (right panel).
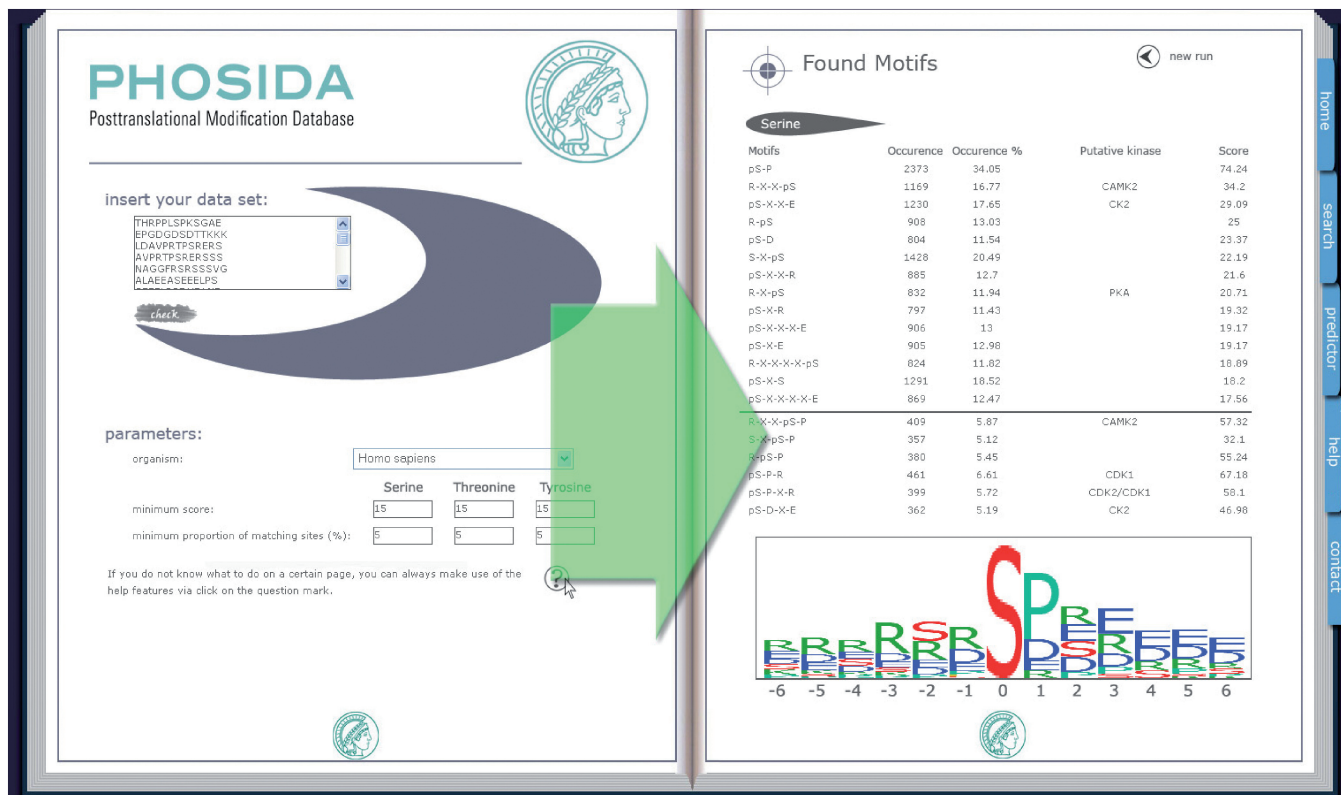

Figure 5. The Motif Finder identifies significantly overrepresented consensus sequences in given large-scale phospho data sets. It compares the position specific amino acid frequencies in the input set (left panel) with the ones of the background set. Based on bootstrap statistics the motif finder extracts *de novo* sequence motifs and matches them with annotated kinase motifs (right panel).

motif finder are available via the help section of PHOSIDA.

FUTURE PLANS AND CONCLUSIONS

As a dynamic database, PHOSIDA is continuously extended and upgraded. The integration of multiple PTM data sets from various species and cell types along with various annotation data makes PHOSIDA a unique environment. In particular the provision of analysis tools and predictors proved to be very helpful. The feedback of the scientific community has increased the usefulness of PHOSIDA consistently over the last four years. While further large-scale data sets will be integrated in the future, we intend to expand the toolkit section. We aim to implement additional analysis tools and extend the motif finder to identify *de novo* motifs for any PTM. The 'news' section provides information about changes and updates on a regular basis. Moreover, users can subscribe to a newsletter.

REFERENCES

1. Hunter,T. (2000) Signaling–2000 and beyond. *Cell*, **100**, 113–127.
2. Pawson,T. and Nash,P. (2003) Assembly of cell regulatory systems through protein interaction domains. *Science*, **300**, 445–452.
3. Choudhary,C. and Mann,M. (2010) Decoding signalling networks by mass spectrometry-based proteomics. *Nature Reviews*, **11**, 427–439.
4. Olsen,J.V., Blagoev,B., Gnad,F., Macek,B., Kumar,C., Mortensen,P. and Mann,M. (2006) Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell*, **127**, 635–648.
5. Daub,H., Olsen,J.V., Bairlein,M., Gnad,F., Oppermann,F.S., Korner,R., Greff,Z., Keri,G., Stemmann,O. and Mann,M. (2008) Kinase-selective enrichment enables quantitative phosphoproteomics of the kinome across the cell cycle. *Mol. Cell*, **31**, 438–448.
6. Pan,C., Gnad,F., Olsen,J.V. and Mann,M. (2008) Quantitative phosphoproteome analysis of a mouse liver cell line reveals specificity of phosphatase inhibitors. *Proteomics*, **8**, 4534–4546.
7. Olsen,J.V., Vermeulen,M., Santamaria,A., Kumar,C., Miller,M.L., Jensen,L.J., Gnad,F., Cox,J., Jensen,T.S., Nigg,E.A. *et al.* (2010) Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. *Sci. Signal*, **3**, ra3.
8. Gnad,F., Ren,S., Cox,J., Olsen,J.V., Macek,B., Oroshi,M. and Mann,M. (2007) PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol.*, **8**, R250.
9. Choudhary,C., Kumar,C., Gnad,F., Nielsen,M.L., Rehman,M., Walther,T.C., Olsen,J.V. and Mann,M. (2009) Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science*, **325**, 834–840.
10. Zielinska,D.F., Gnad,F., Wisniewski,J.R. and Mann,M. Precision mapping of an in vivo N-glycoproteome reveals rigid topological and sequence constraints. *Cell*, **141**, 897–907.
11. The UniProt Consortium. (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.*, **38**, D142–D148.
12. Diella,F., Gould,C.M., Chica,C., Via,A. and Gibson,T.J. (2008) Phospho.ELM: a database of phosphorylation sites–update 2008. *Nucleic Acids Res.*, **36**, D240–D244.
13. Hornbeck,P.V., Chabra,I., Kornhauser,J.M., Skrzypek,E. and Zhang,B. (2004) PhosphoSite: a bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics*, **4**, 1551–1561.
14. Keshava Prasad,T.S., Goel,R., Kandasamy,K., Keerthikumar,S., Kumar,S., Mathivanan,S., Telikicherla,D., Raju,R., Shafreen,B., Venugopal,A. *et al.* (2009) Human Protein Reference Database–2009 update. *Nucleic Acids Res.*, **37**, D767–D772.
15. Lee,T.Y., Huang,H.D., Hung,J.H., Huang,H.Y., Yang,Y.S. and Wang,T.H. (2006) dbPTM: an information repository of protein post-translational modification. *Nucleic Acids Res.*, **34**, D622–D627.
16. Macek,B., Mann,M. and Olsen,J.V. (2009) Global and site-specific quantitative phosphoproteomics: principles and applications. *Annu. Rev. Pharmacol. Toxicol.*, **49**, 199–221.
17. Gnad,F., Forner,F., Zielinska,D.F., Birney,E., Gunawardena,J. and Mann,M. (2010) Evolutionary constraints of phosphorylation in eukaryotes, prokaryotes and mitochondria. *Mol. Cell Proteomics*, [Epub ahead of print, 5 August 2010].
18. Rinehart,J., Maksimova,Y.D., Tanis,J.E., Stone,K.L., Hodson,C.A., Zhang,J., Risinger,M., Pan,W., Wu,D., Colangelo,C.M. *et al.* (2009) Sites of regulated phosphorylation that control K-Cl cotransporter activity. *Cell*, **138**, 525–536.
19. Seidler,J., Adal,M., Kubler,D., Bossemeyer,D. and Lehmann,W.D. (2009) Analysis of autophosphorylation sites in the recombinant catalytic subunit alpha of cAMP-dependent kinase by nano-UPLC-ESI-MS/MS. *Anal. Bioanal. Chem.*, **395**, 1713–1720.
20. Matic,I., Schimmel,J., Hendriks,I.A., van Santen,M.A., van de Rijke,F., van Dam,H., Gnad,F., Mann,M. and Vertegaal,A.C. (2010) Site-specific identification of SUMO-2 targets in cells reveals an inverted SUMOylation motif and a hydrophobic cluster SUMOylation motif. *Mol. Cell*, **39**, 641–652.
21. Oppermann,F.S., Gnad,F., Olsen,J.V., Hornberger,R., Greff,Z., Keri,G., Mann,M. and Daub,H. (2009) Large-scale proteomics analysis of the human kinome. *Mol. Cell Proteomics*, **8**, 1751–1764.
22. Zanivan,S., Gnad,F., Wickstrom,S.A., Geiger,T., Macek,B., Cox,J., Fassler,R. and Mann,M. (2008) Solid tumor proteome and phosphoproteome analysis by high resolution mass spectrometry. *J. Proteome Res.*, **7**, 5314–5326.
23. Wisniewski,J.R., Nagaraj,N., Zougman,A., Gnad,F. and Mann,M. (2010) Brain phosphoproteome obtained by a FASP-based method reveals plasma membrane protein topology. *J. Proteome Res.*, **9**, 3280–3289.
24. Weintz,G., Olsen,J.V., Fruhauf,K., Niedzielska,M., Amit,I., Jantsch,J., Mages,J., Frech,C., Dolken,L., Mann,M. *et al.* (2010) The phosphoproteome of toll-like receptor-activated macrophages. *Mol. Syst. Biol.*, **6**, 371.
25. Hilger,M., Bonaldi,T., Gnad,F. and Mann,M. (2009) Systems-wide analysis of a phosphatase knock-down by quantitative proteomics and phosphoproteomics. *Mol. Cell Proteomics*, **8**, 1908–1920.
26. Zielinska,D.F., Gnad,F., Jedrusik-Bode,M., Wisniewski,J.R. and Mann,M. (2009) Caenorhabditis elegans has a phosphoproteome atypical for metazoans that is enriched in developmental and sex determination proteins. *J. Proteome Res.*, **8**, 4039–4049.
27. Gnad,F., de Godoy,L.M., Cox,J., Neuhauser,N., Ren,S., Olsen,J.V. and Mann,M. (2009) High-accuracy identification and

bioinformatic analysis of in vivo protein phosphorylation sites in yeast. *Proteomics*, **9**, 4642–4652.

28. Macek,B., Gnad,F., Soufi,B., Kumar,C., Olsen,J.V., Mijakovic,I. and Mann,M. (2008) Phosphoproteome analysis of E. coli reveals evolutionary conservation of bacterial Ser/Thr/Tyr phosphorylation. *Mol. Cell Proteomics*, **7**, 299–307.

29. Macek,B., Mijakovic,I., Olsen,J.V., Gnad,F., Kumar,C., Jensen,P.R. and Mann,M. (2007) The serine/threonine/tyrosine phosphoproteome of the model bacterium Bacillus subtilis. *Mol. Cell Proteomics*, **6**, 697–707.

30. Aivaliotis,M., Macek,B., Gnad,F., Reichelt,P., Mann,M. and Oesterhelt,D. (2009) Ser/Thr/Tyr protein phosphorylation in the archaeon Halobacterium salinarum–a representative of the third domain of life. *PLoS ONE*, **4**, e4777.

31. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, **25**, 25–29.

32. Carbon,S., Ireland,A., Mungall,C.J., Shu,S., Marshall,B. and Lewis,S. (2009) AmiGO: online access to ontology and annotation data. *Bioinformatics*, **25**, 288–289.

33. Bradshaw,R.A., Medzihradszky,K.F. and Chalkley,R.J. (2010) Protein PTMs: post-translational modifications or pesky trouble makers? *J. Mass Spectrom.*, **45**, 1095–1097.

34. Flicek,P., Aken,B.L., Ballester,B., Beal,K., Bragin,E., Brent,S., Chen,Y., Clapham,P., Coates,G., Fairley,S. *et al.* (2010) Ensembl's 10th year. *Nucleic Acids Res.*, **38**, D557–562.

35. Obenauer,J.C., Cantley,L.C. and Yaffe,M.B. (2003) Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res.*, **31**, 3635–3641.

36. Blom,N., Gammeltoft,S. and Brunak,S. (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.*, **294**, 1351–1362.

37. Blom,N., Sicheritz-Ponten,T., Gupta,R., Gammeltoft,S. and Brunak,S. (2004) Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics*, **4**, 1633–1649.

38. Durek,P., Schudoma,C., Weckwerth,W., Selbig,J. and Walther,D. (2009) Detection and characterization of 3D-signature phosphorylation site motifs and their contribution towards improved phosphorylation site prediction in proteins. *BMC Bioinformatics*, **10**, 117.

39. Biswas,A.K., Noman,N. and Sikder,A.R. (2010) Machine learning approach to predict protein phosphorylation sites by incorporating evolutionary information. *BMC Bioinformatics*, **11**, 273.

40. Gnad,F., Ren,S., Choudhary,C., Cox,J. and Mann,M. (2010) Predicting post-translational lysine acetylation using support vector machines. *Bioinformatics*, **26**, 1666–1668.

41. Li,S., Li,H., Li,M., Shyr,Y., Xie,L. and Li,Y. (2009) Improved prediction of lysine acetylation by support vector machines. *Protein Pept. Lett.*, **16**, 977–983.

42. Basu,A., Rose,K.L., Zhang,J., Beavis,R.C., Ueberheide,B., Garcia,B.A., Chait,B., Zhao,Y., Hunt,D.F., Segal,E. *et al.* (2009) Proteome-wide prediction of acetylation substrates. *Proc. Natl Acad. Sci. USA*, **106**, 13785–13790.